Project 1 -- Environment

Think of an agent that plays a 2-armed bandit, trying to maximize its total reward. In each step, the agent selects one of the levers and is given some reward according to the reward distribution of that lever. Assume that reward distribution for the first lever is a Gaussian with $\mu_1 = 5, \sigma_1^2 = 10$, and for the second lever is a binomial Gaussian with $\mu_{21} = 10, \sigma_{21}^2 = 15, \mu_{22} = 4, \sigma_{22}^2 = 10$, which means that the resulting output will be uniformly probable from these two Gaussian distributions (See http://en.wikipedia.org/wiki/Mixture_distribution).

Implement this environment in Python together with a random policy that chooses the two actions with equal probabilities. Plot the resulting average reward per timestep obtained by the agent up to timestep k for $k = 1, \ldots, 1000$.